Journal of Athletic Training 2006;41(1):93–101

www.journalofathletictraining.org

original research

Reliability of First Ray Position and Mobility Measurements in Experienced and Inexperienced Examiners

Crystal Shirk*; Michelle A. Sandrey†; Mia Erickson†

*****Summersville Memorial Hospital, Summersville, WV; †West Virginia University, Morgantown, WV

Crystal Shirk, MS, ATC; Michelle A. Sandrey, PhD, ATC; and Mia Erickson, EdD, PT, ATC, contributed to conception and design; acquisition and analysis and interpretation of the data; and drafting, critical revision, and final approval of the article. Address correspondence to Crystal Shirk, MS, ATC, 25 Timber Oaks Drive, Craigsville, WV 26205. Address e-mail to clshirk@juno.com.

Context: Neither reliability nor validity data exist for the Root method of clinically assessing first ray position or mobility by experienced and inexperienced examiners.

Objective: To determine intrarater and interrater reliability for first ray position and mobility measurements in experienced and inexperienced examiners.

Design: Single-blind prospective reliability study.

Setting: Physical therapy clinic.

Patients or Other Participants: Four examiners, 2 experienced and 2 inexperienced, obtained first ray position and mobility measurements. Both feet of 36 subjects (14 males, 22 females) were measured.

Intervention(s): Each examiner evaluated first ray position and mobility for each of the subjects' feet on 2 separate occasions using the manual assessment techniques described by Root.

Main Outcome Measure(s): First ray position (normal, plantar flexed, dorsiflexed) and mobility (normal, hypermobile, hypomobile) decisions were made.

Results: We calculated kappa correlation coefficients for intrarater and interrater reliability. For position, intrarater and interrater reliability ranged from .03 to .27 for all examiners, experienced and inexperienced. For mobility, intrarater and interrater reliability ranged from .02 to .26 for experienced, inexperienced, and experienced/inexperienced. The percentage agreement (P_O) values for all examiners were less than 58%. For individual values for position, intrarater and interrater reliability ranged from .00 to .26. For individual values for mobility, intrarater and interrater reliability ranged from .00 to .26. The P_O values for all examiners were less than 50%.

Conclusions: Clinical experience was not associated with higher kappa coefficients or P_O values when examiners assessed first ray position or mobility. Clinicians should acknowledge the poor reliability of first ray measurements, especially when making treatment decisions. Finally, a validity study to compare the Root techniques with a gold standard is warranted.

Key Words: foot biomechanics, foot examination, orthotics

The first ray consists of the first metatarsal and first cuneiform[1–4] and serves important purposes during the gait cycle: providing shock absorption during the loading response and stability during the terminal stance and push-off phases of the gait cycle. Abnormal first ray position (plantar flexion or dorsiflexion) or abnormal mobility (hypermobility or hypomobility) decreases the structure's ability to function normally during gait.[2] First ray abnormalities have been suggested as a causative factor for the development of metatarsalgia.[5] Experimentally, associations have been found between first ray abnormalities and hallux valgus,[6,7] forefoot valgus,[2] rheumatoid acquired flatfoot,[8] and plantar ulcerations.[9] In addition, abnormal first ray mobility has also been highly correlated with excessive knee rotation and altered ground reaction forces during gait.[10]

First ray position and mobility are often included as part of a biomechanical examination, and orthotic modifications are often made for individuals with first ray abnormalities (ie, first ray cut out). Because of the relationship among abnormal first ray mechanics, lower extremity abnormalities, and orthotic intervention, first ray assessment is an important aspect of the lower extremity examination.

Examination of the first ray's position and mobility can be performed using radiographs[4,11] or a first ray mobility measuring device.[12,13] Glasoe et al[12] reported both high reliability (.98) and high validity (.97) for the first ray mobility measuring device when using radiographs as the gold standard. Although the measuring device was reported to be valid when compared with radiographs, neither the device nor radiographs are readily available or practical in a sports medicine setting.

Clinically, manual methods are used for assessment of first ray position and mobility. Root et al[3] suggested one method for clinically assessing first ray position, and Root et al[3,4] and Glasoe et al[13] suggested techniques for assessing mobility. Glasoe et al[13] found moderate to substantial intrarater (test-retest) reliability (.50 to .85) but slight interrater reliability (.09

to .16) for both experienced and inexperienced examiners using their technique. Validity of the Glasoe et al manual technique was poor (−.21) when compared with findings from a first ray mobility measuring device.[13] Cornwall et al[14] also found that the Glasoe et al manual technique for measuring first ray mobility had poor interrater reliability (.01 to .20) among 3 clinicians with 6 or more years of experience. Validity was also poor when manual methods were compared with measurements from a similar first ray measuring device (.01 to .30).[14]

Low interrater reliability coefficients among experienced and inexperienced examiners led us to question the clinical value of the Glasoe et al technique for measuring first ray mobility as well as the role of experience when performing this clinical assessment. Neither reliability nor validity data exist for the assessment techniques (position or mobility) described by Root et al.[3] Therefore, our purposes were to determine intrarater (test-retest) and interrater reliability for first ray position and mobility assessment techniques as described by Root et al[3,4] for experienced and inexperienced examiners. Intrarater reliability coefficients were expected to exceed those for interrater reliability. In addition, experienced examiners were expected to demonstrate higher intrarater and interrater reliability coefficients than inexperienced examiners.

METHODS

Research Design

The research design was a single-blind prospective reliability study in which we examined intrarater (test-retest) reliability of first ray position (plantar flexed, dorsiflexed, or normal) and mobility (hypermobile, hypomobile, or normal) assessments for experienced and inexperienced examiners. Intrarater reliability was determined for both experienced and inexperienced examiners. Intrarater reliability was also determined between experienced and inexperienced examiners.

Subjects

Examiners and Recorders. Four examiners, 2 experienced (men) and 2 inexperienced (women) were recruited for the study. *Experienced* was defined as a certified athletic trainer and/or licensed physical therapist with 6 or more years of clinical experience who routinely works with patients having lower extremity dysfunction. Experienced testers included 1 physical therapist/certified athletic trainer (E-1) who works in a physical therapy clinic and 1 certified athletic trainer (E-2) who works in a university athletic training setting. *Inexperienced* was defined as a certified athletic trainer and/or licensed physical therapist with less than 2 years of clinical experience. Two certified athletic trainers (I-1 and I-2) who were second-year graduate athletic training students and had completed classes in anatomy and lower extremity biomechanics served as inexperienced examiners. Five additional individuals were recruited to record all data for both days. Examiners were blinded to all previous measurements and subject identities.

Subjects. Thirty-six subjects (14 males, 22 females; average age, 23 ± 5.93 years) were recruited from a convenience sample of the local college student population. Each subject volunteered both feet (n = 72 feet). The only exclusion criterion for these subjects was a history of foot surgery.[13] All examiners, recorders, and subjects signed an informed consent form approved by the University's Institutional Review Board for the Protection of Human Subjects, which also approved the study. Only the subjects completed demographic and medical history questionnaires before participating.

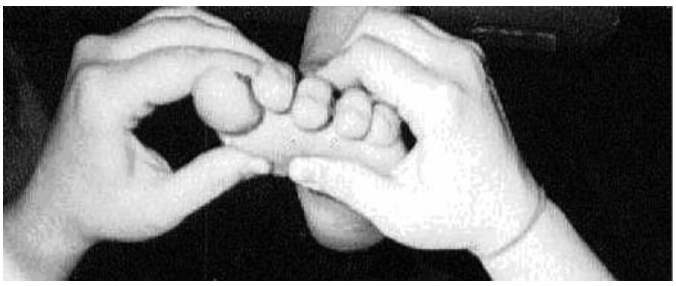

Figure 1. View of the lumbrical grip used while testing first ray position, which was graded as normal.

Evaluation Protocol

Before the first testing session, all examiners were shown the manual method for assessing first ray position and mobility as described by Root et al.[3,4] The examiners were given 45 minutes to practice the 2 testing procedures. They were instructed to evaluate first ray position and mobility with the subtalar joint in a neutral position. For this study, subtalar joint neutral was determined by placing the thumb and forefinger on either side of the talar dome. The subtalar joint was then pronated and supinated until the talus could be felt equally on both sides. In addition, examiners were instructed to use a comfortable amount of skin pressure for both position and mobility testing, which displaced the skin enough by palpation that the metatarsal heads were felt.

First Ray Position. The position of the first ray was determined by how it lies in comparison with the lateral 4 metatarsals.[1] The examiner grasped the plantar and dorsal aspects of the first metatarsal head between the pad of one thumb and the corresponding index finger. The lateral 4 metatarsal heads were grasped between the thumb and remaining digits of the opposite hand (lumbrical grip) (Figure 1). Examiners were instructed to use pressure to lightly compress the plantar fat pads to palpate the metatarsal heads. If the first metatarsal head lay in the same plane as the remaining 4, it was graded as normal (see Figure 1). If the first metatarsal head lay above (dorsal to) the remaining 4, it was graded as dorsiflexed (Figure 2). If the first metatarsal head lay below the remaining 4, it was graded as plantar flexed (Figure 3).

First Ray Mobility. To measure first ray mobility, the examiner grasped the metatarsal heads as described for position testing above. Using the lumbrical grip, the examiner stabilized metatarsal heads 2 through 5 and displaced the first metatarsal in the dorsal and plantar directions until a capsular endpoint was felt. When the amount of dorsal movement (dorsiflexion) exceeded the amount of plantar movement (plantar flexion), the first ray was graded as hypermobile. When the amount of plantar flexion exceeded the amount of dorsiflexion, the first ray was graded as hypomobile. When the amounts of plantar flexion and dorsiflexion were equal, the ray was graded as normal.

Figure 2. First ray position graded as dorsiflexed.

Figure 3. First ray position graded as plantar flexed.

Figure 4. Subject covered by a curtain draped over the lower legs.

Testing Procedures

Subjects were tested in groups of 4 on 2 separate occasions. When subjects arrived, they were escorted to the testing area by the principal investigator. Tables were placed 10 ft (3.05 m) apart and were separated by a curtain draped around each subject. The end of the curtain fell across the subject's lower legs, so that only the subject's feet were visible to the examiner (Figure 4). Examiners waited in a closed room while the primary investigator positioned the subjects.

Once subjects were positioned appropriately, the 4 examiners entered the room; then each examiner positioned himself or herself at the right foot of 1 of the subjects, so that all subjects were being evaluated simultaneously. The examiners evaluated first ray position and mobility as described above. To ensure that the correct foot was measured, the recorder placed a towel over the opposite foot. Once the examiner determined position and mobility grades, he or she quietly reported them to the designated recorder, who subsequently recorded all grades for position and mobility on a data sheet. Examiners were not able to access findings once they were reported to the recorder. When each examiner was finished, he or she rotated to the right foot of the next subject. This process was repeated until all right feet were examined. Each subject's right foot was measured before the left to decrease the chance of the examiner's recalling right foot measurements and, thus, biasing the decision regarding the left foot. Once all right feet were examined and each examiner was back to his or her initial subject, the process continued with the subjects' left feet until both feet of all subjects were tested by each examiner. Each recorder rotated with his or her assigned examiner during each testing session and recorded all data for the examiner. At the end of the day, the completed data sheet was given to the principal investigator.

Test-Retest Procedures

To complete the intrarater (test-retest) portion of this study, all subjects returned for a second evaluation 7 days later. The examiners, subjects, and recorders followed the same testing procedure on day 2 as on day 1. Recorders used a new data sheet on the second day of testing.

Blind coding kept the examination results from both sessions anonymous and allowed subject data to be compared for the intrarater portion of the study. Each subject was assigned a code (A1, A2, etc) by the primary investigator. This code was placed on the subject's data sheets for testing sessions 1 and 2. After the primary investigator positioned the subject on the table for both testing sessions, the subject's code was placed on the subject's table. This allowed the primary investigator to pair data for each subject from both testing sessions accurately. Only the primary investigator had access to the master list of codes.

Statistical Analyses

We calculated means and standard deviations to describe subjects' demographic information. Assessment techniques required the examiner to classify first ray position as plantar flexed, dorsiflexed, or normal and mobility as hypermobile, hypomobile, or normal. Using these techniques, no numeric data were produced; rather the examiner made a clinical judgment to assign a label for position and mobility. Therefore, data in this study were nominal, and kappa reliability (κ) coefficients and percentage agreement (P_O) values were calcu-

Table 1. Intrarater and Interrater Reliability of First Ray Position Measurements for All Examiners*

	κ	*P*	P_O
Intrarater Reliability			
Examiners			
All	.24	.001†	49.7
Experienced	.21	.001†	45.7
Inexperienced	.27	.001†	57.5
Interrater Reliability			
Examiners			
All	.03	.653	34.5
Experienced	.12	.006‡	42.6
Inexperienced	.11	.002‡	38.6

*κ indicates kappa coefficient; P_O, percent agreement.
†$P \leq .001$.
‡$P \leq .01$.

Table 2. Intrarater Reliability of First Ray Position for Individual Examiners*

Examiner	P_O	κ (95% CI)	*P*	κ_{max}	κ/κ_{max} (%)
E-1	50	.26	.001†	.65	40.3
E-2	42	.13	.098	.62	20.6
I-1	67	.00	.930	.28	00.0
I-2	49	.24	.002†	.55	42.7

*P_O indicates percent agreement; κ, kappa coefficient; CI, confidence interval; κ_{max}, maximum kappa coefficient; E, experienced; I, inexperienced.
†$P \leq .01$.

Table 3. Interrater Reliability of First Ray Position for Individual Examiners

Examiner	P_O	κ (95% CI)	*P*	κ_{max}	κ/κ_{max} (%)
Session 1					
E-1 × E-2	43	.11	.188	.69	15.2
I-2 × E-2	40	.14	.062	.36	38.6
I-2 × E-1	39	.10	.203	.56	17.3
I-1 × E-2	22	.02	.558	.11	18.4
I-1 × E-1	43	.01	.875	.16	4.4
I-1 × I-2	44	.03	.553	.22	14.5
Session 2					
E-1 × E-2	26	.00	.148	.77	0.0
I-2 × E-2	51	.18	.035	.86	20.9
I-2 × E-1	40	.09	.267	.35	24.8
I-1 × E-2	32	.01	.821	.32	4.3
I-1 × E-1	32	.00	.695	.39	0.0
I-1 × I-2	42	.14	.025	.30	47.3

*P_O indicates percent agreement; κ, kappa coefficient; CI, confidence interval; κ_{max}, maximum kappa coefficient; E, experienced; I, inexperienced.

lated.[15,16] Intrarater reliability κ coefficients and P_O values were determined for first ray position and mobility for (1) all examiners, (2) experienced and inexperienced examiners, and (3) individual examiners. Intrarater reliability was determined by comparing each examiner's results between the 2 testing sessions (test-retest). Interrater reliability κ coefficients and P_O values were determined for first ray position and mobility for (1) experienced and inexperienced examiners and (2) individual examiners for sessions 1 and 2.

To interpret κ coefficients, the following scale by Landis and Koch[17] was used[18]: .81 to 1.0 = almost perfect agreement, .61 to .80 = substantial agreement, .41 to .60 = moderate agreement, .21 to .40 = fair agreement, .00 to .20 = slight agreement, and <.00 = poor agreement. The alpha level for κ coefficients was set at $P < .01$. We used SPSS for Windows (version 11.0.0; SPSS Inc, Chicago, IL) for determining κ coefficients and P_O values.

Kappa correlation coefficients adjust for agreement that has occurred by chance when there have been either (1) a small number of nominal categories (as is the case in this study, ie, plantar flexed, dorsiflexed, and normal) or (2) uneven data distributions (which occur when an examiner has chosen 1 of the categories more frequently than the others). Uneven data distributions could cause κ coefficients to be low even when P_O values are high.[15,19] To estimate the effect of uneven data distributions in this study, we calculated the maximum kappa (κ_{max}) correlation coefficients for individual examiners. The κ_{max} coefficients represent the best reliability that could have occurred for this data distribution. In addition, κ/κ_{max} values were calculated to determine the proportion of the κ_{max} that each examiner was able to reach. High κ/κ_{max} values indicate that 2 examiners could agree well, even though the data distribution produced low κ coefficients.[19] Maximum kappa correlation coefficients and κ/κ_{max} values were calculated using the formula provided by Cohen.[15]

RESULTS

Position

Intrarater reliability κ coefficients and P_O values for first ray position for all examiners exceeded those for interrater reliability (Table 1). Intrarater κ coefficients ranged from .21 to .27 (fair agreement), and P_O values from 45.7% to 57.5%. Interrater κ coefficients ranged from .03 to .12 (slight agreement), and P_O values ranged from 34.5% to 42.6%. Inexperienced examiners' κ coefficients and P_O values exceeded those of experienced examiners for both intrarater and interrater reliability. All intrarater and interrater κ coefficients reached statistical significance ($P < .006$) except the interrater κ coefficient for all examiners.

Kappa intrarater reliability coefficients for individual examiners ranged from .00 to .26 (poor to fair agreement), and P_O values ranged from 42% (E-2) to 67% (I-1) (Table 2). For intrarater reliability, I-1 had the lowest κ coefficient but the best P_O value. The intrarater κ_{max} coefficients ranged from .28 (I-1) to .65 (E-1), and intrarater κ/κ_{max} values ranged from 0.0% (I-1) to 42.7% (I-2). The only statistically significant κ coefficients existed for examiners E-1 ($\kappa = .26$, $P = .001$) and I-2 ($\kappa = .24$, $P = .002$).

Kappa interrater reliability coefficients between individual examiners for first ray position ranged from .00 to .18 (poor to slight agreement) (Table 3). P_O values ranged from 22% (I-1 and E-2) to 44% (I-1 and I-2) for session 1 and from 26% (E-1 and E-2) to 51% (I-2 and E-2) for session 2. Interrater κ_{max} coefficients ranged from .11 (I-1 and E-2) to .69 (E-1 and E-2) for session 1 and from .30 (I-1 and I-2) to .86 (I-2

Table 4. Intrarater and Interrater Reliability of First Ray Mobility for All Examiners*

	κ	*P*	P_O
Intrarater Reliability			
Examiners			
All	.16	.001†	44.1
Experienced	.03	.628	36.8
Inexperienced	.26	.001†	51.4
Interrater Reliability			
Examiners			
All	.02	.689	35.8
Experienced	.12	.005‡	43.9
Inexperienced	.14	.001‡	42.2

*κ indicates kappa coefficient; P_O, percent agreement.
†$P \leq .001$.
‡$P \leq .01$.

Table 5. Intrarater Reliability of First Ray Mobility for Individual Examiners*

Examiner	P_O	κ (95% CI)	*P* Value	κ_{max}	κ/κ_{max} (%)
E-1	40	.07	.403	.74	9.3
E-2	52	.00	.838	.83	0.0
I-1	57	.26	.001†	.52	50.6
I-2	46	.18	.010‡	.48	37.7

*P_O indicates percent agreement; κ, kappa coefficient; CI, confidence interval; κ_{max}, maximum kappa coefficient; E, experienced; I, inexperienced.
†$P \leq .001$.
‡$P \leq .01$.

and E-2) for session 2. Interrater κ/κ_{max} values ranged from 4.4% (I-1 and E-1) to 38.6% (I-2 and E-2) for session 1 and from 0.0% (I-1 and E-1; E-2 and E-1) to 47.3% (I-1 and I-2) for session 2. No statistically significant κ coefficients were noted for interrater reliability for position measurements during either testing session.

Mobility

Intrarater κ coefficients for first ray mobility ranged from .03 to .26 (slight to fair agreement), and P_O values ranged from 36.8% to 51.4% (Table 4). Interrater κ coefficients ranged from .02 to .14 (slight agreement), and P_O values ranged from 35.8% to 43.9%. For mobility, κ coefficients for intrarater and interrater reliability were similar except for the intrarater κ coefficient for the inexperienced examiners ($\kappa = .26$), which did reach the low margin of the fair category. For mobility, inexperienced examiners' κ coefficients and P_O values exceeded those of the experienced examiners. All intrarater and interrater κ coefficients reached statistical significance ($P < .005$) except for those of the experienced examiners.

Intrarater reliability coefficients for individual examiners for first ray mobility ranged from .00 to .26 (poor to fair agreement), and P_O values ranged from 40% (E-1) to 57% (I-1) (Table 5). The intrarater κ_{max} coefficients ranged from .48 (I-2) to .83 (E-2), and intrarater κ/κ_{max} values ranged from 0.0% (E-2) to 50.6% (I-1). The only statistically significant intrarater

Table 6. Interrater Reliability of First Ray Mobility for Individual Examiners

Examiner*	P_O	κ (95% CI)	*P*	κ_{max}	κ/κ_{max} (%)
Session 1					
E-1 × E-2	39	.09	.218	.63	14.6
I-2 × E-2	38	.00	.966	.82	0.5
I-2 × E-1	38	.11	.106	.58	18.1
I-1 × E-2	44	.08	.288	.49	16.3
I-1 × E-1	40	.02	.719	.41	5.9
I-1 × I-2	46	.05	.502	.54	9.8
Session 2					
E-1 × E-2	35	.01	.907	.85	1.2
I-2 × E-2	24	.00	.304	.49	0.0
I-2 × E-1	46	.22	.003†	.49	44.1
I-1 × E-2	35	.11	.178	.67	16.3
I-1 × E-1	42	.04	.644	.77	4.7
I-1 × I-2	42	.13	.061	.51	26.1

*P_O indicates percent agreement; κ, kappa coefficient; CI, confidence interval; κ_{max}, maximum kappa coefficient; E, experienced; I, inexperienced.
†$P \leq .01$.

coefficients for mobility existed for examiners I-1 ($\kappa = .26$, $P = .001$) and I-2 ($\kappa = .18$, $P = .010$).

Kappa interrater reliability coefficients between examiners for first ray mobility ranged from .00 to .22 (poor to fair agreement) (Table 6). The P_O values ranged from 38% (I-2 and E-2) to 46% (I-1 and I-2) for session 1 and from 24% (I-2 and E-2) to 46% (I-2 and E-1) for session 2. Interrater κ_{max} coefficients ranged from .41 (I-1 and E-1) to .82 (I-2 and E-2) for session 1 and from .49 (I-2 and E-1; I-2 and E-2) to .85 (E-1 and E-2) for session 2. Interrater κ/κ_{max} values ranged from 0.5% (I-2 and E-2) to 18.1% (I-2 and E-1) for session 1 and from 0.0% (I-2 and E-2) to 44.1% (I-2 and E-1) for session 2. The only statistically significant result existed between I-2 and E-1 in session 2 ($\kappa = .22$, $P = .003$).

DISCUSSION

For all examiners, intrarater reliability for position testing reached the low margin of the fair category, whereas intrarater reliability for mobility was only slight. Examiners agreed on classification in 49.7% of the subjects for position and 44.1% for mobility. We hypothesized that intrarater reliability would exceed interrater reliability for both position and mobility testing. For position testing, overall intrarater κ coefficients reached the low margin of the fair category, whereas interrater coefficients were only slight. For mobility, overall intrarater and interrater coefficients were similar (slight) except for the intrarater κ coefficients for the inexperienced examiners, which again reached the low margin of the fair category (see Table 1).

We hypothesized that experienced examiners would have better intrarater and interrater reliability than inexperienced examiners; however, this was not the case. Intrarater and interrater κ coefficients and P_O values for the inexperienced examiners exceeded those of the experienced examiners for both position and mobility. These results suggest that clinical experience was not associated with higher κ coefficients or P_O values when examining first ray position or mobility.

For individual examiners, the highest intrarater κ coefficients reached the low margin of the fair category for E-1 (κ = .26, intrarater position), I-1 (κ = .026, intrarater mobility), and I-2 (κ = .24, intrarater position). These κ coefficients corresponded to P_O values of 50%, 57%, and 49%, respectively. These 3 examiners also demonstrated the highest κ/κ_{max} values (40.3%, 50.6%, and 42.7%, respectively). This suggests these κ coefficients might have been influenced by uneven data distributions. Uneven data distributions will cause 2 raters having high agreement to have low κ coefficients.[15,17,19] The highest κ/κ_{max} value was 50.6% for intrarater reliability (mobility) for I-1, who had the tendency to choose normal for mobility (59%). This inexperienced examiner may have been biased to choose normal based on the sample of healthy participants in the study. Inexperience may have influenced the classification of subjects by examiner I-1.

The highest intrarater κ_{max} coefficient (E-2, κ = .83, mobility) corresponded with a low κ coefficient (κ = .00) but a P_O value of 52% (similar to E-1, I-1, and I-2, whose κ coefficients reached the fair category). For this examiner, the κ/κ_{max} was 0.00%. In addition, the examiner with the highest P_O value (I-1, P_O = 67%, position) also obtained a low κ coefficient and κ/κ_{max} value (κ = .00, κ/κ_{max} = 0.0%). This latter examiner, I-1, had a tendency to choose *normal* (81% of subjects), but the κ_{max} coefficient was .28. Therefore, the uneven data distribution did not affect the κ coefficient, but inexperience and the sample of healthy subjects may have again influenced I-1 to classify more subjects as normal.

A limitation to this study may be that participants were normal, healthy individuals aged 18 to 39 years. Other authors[20] have indicated that biomechanical abnormalities are present within a healthy population of subjects. In addition, authors[14] of a recent study of 30 healthy subjects (both feet, n = 60) to examine the first ray indicated that their sample was representative of the normal population. Nevertheless, examiners might have been better able to reproduce findings in a symptomatic population. Also, the potential for examiner bias would have been eliminated. It is important to point out that the only exclusion criterion was a history of foot surgery. Individuals with past or present lower extremity abnormalities alone were not excluded.

Ours is the first study to examine intrarater and interrater reliability of the first ray position and mobility measurement techniques described by Root et al.[3,4] Our findings suggest low intrarater reliability for position and mobility measurements, regardless of the examiner's experience using the Root et al techniques. Other authors have studied the Glasoe et al[13] technique for measuring first ray mobility, with different results for intrarater reliability. Glasoe et al[13] indicated moderate to substantial intrarater reliability using this technique. The Glasoe et al[13] method requires the examiner to apply a dorsal force on the first metatarsal head and compare movement with the position of the lateral 4 metatarsals, whereas the Root et al[3,4] technique compares mobility of the first ray when both dorsal and plantar forces are applied. The dorsal and plantar forces in the Root et al techniques could provide an additional source of error. Additionally, the Glasoe et al technique for measuring mobility assumes that the individual's first ray rests in a plantar-flexed position, and a normal mobility grade is assigned if the metatarsal head can be dorsiflexed to the level of the other metatarsal heads. Perhaps it is easier for an examiner to determine mobility in these individuals using the Glasoe et al grading system, thus providing better intrarater reliability.

Interrater reliability κ coefficients for individual examiners yielded only 1 κ coefficient within the fair category (κ = .22, I-2 × E-1, mobility, session 2). The P_O values did not exceed 51% between examiners for position or mobility. Interrater reliability coefficients found in this study are consistent with other studies using the Glasoe et al technique.[13,14] This suggests low interrater reliability in both previously reported methods to evaluate first ray mobility.

Other possible reasons for low reliability coefficients include insufficient practice time and inadequate standardization of force application and foot position. All examiners were given an opportunity to review and practice before the second testing session, but all felt comfortable from the previous week and declined additional practice. Because P_O values for sessions 1 and 2 were similar for all examiners, no practice effect was apparent.

The pressure level used when dorsiflexing and plantar flexing the first ray might have been different for each individual. One examiner may not have applied as much pressure as another because of inexperience, weakness, or apprehension. In a study[11] of a first ray mobility measuring device, when forces of 20, 35, 55, and 85 N were separately applied to the first ray, 55 N produced the best force with the least unwanted movement in the forefoot and rearfoot. A limitation of our study was that force application was not standardized, and this could have caused inconsistencies in evaluating first ray mobility. This lack of standardization could also be true for a clinical setting, where force application is not typically measured when examining first ray mobility. Standardizing force application could improve consistency among examiners.

Another source of error may be lack of standardization of the subtalar and talocrural joint positions during the first ray examination. The procedure we used required the examiners to place the foot in subtalar joint neutral. Each examiner determined subtalar joint neutral independently. Previous authors [21,22] have indicated that intertester reliability of subtalar joint neutral position is poor when foot position is not standardized, thus introducing another source of error when examining the first ray. Inconsistencies in subtalar positioning may have influenced the position of the first ray, thereby decreasing reliability coefficients in this study. Bevans[23] indicated that examining the first ray with the calcaneus in eversion or inversion (which often occurs if the subtalar joint is not in a neutral position) causes changes in first ray dorsiflexion. For example, when the calcaneus is in eversion, first ray dorsiflexion increases, and when the calcaneus is in inversion, first ray dorsiflexion decreases. It is important to point out that the subtalar joint is maintained in neutral with the first ray measuring device. This is a potential reason for its high reliability when compared with clinical measurement techniques.

Grebing and Coughlin[24] reported increased first ray motion with talocrural joint plantar flexion and decreased motion with talocrural dorsiflexion when compared with a neutral position. Examiners were not required to standardize talocrural joint position. Our results, combined with findings from Bevans[23] and Grebing and Coughlin,[24] indicate that criteria to standardize both subtalar and talocrural joints might be necessary when evaluating the first ray. Future authors should examine the reliability of first ray position and mobility measurements with both subtalar and talocrural joint positions standardized for each examiner.

Manual first ray measurement techniques previously described by Glasoe et al[13] to assess first ray mobility have not been proven valid. A limitation of our study is that examiners' classifications for position and mobility were not compared with a gold standard. This still leaves us to question the validity of the Root et al techniques. More research is needed on the Root et al techniques, in which examiner findings are compared with radiographic or first ray measuring device findings. Valuable data would be provided to clinicians, including sensitivity and specificity of the techniques.

When using correlation coefficients, statistical significance does not imply clinical meaningfulness. Although several κ coefficients were statistically significant, the clinical value of the first ray position and mobility measurements is limited by the low degree of reliability the κ coefficients represent. Poor reliability raises questions about the utility of these assessment techniques, particularly in relation to their use as the basis for clinical decisions. Although some clinicians consider first ray assessment to be an important component of lower extremity evaluation,[10,25] our results suggest that improved clinical techniques for categorizing first ray position and mobility are needed for accurate assessment of a patient's status.

CONCLUSIONS

Both experienced and inexperienced examiners demonstrated low reliability when measuring first ray position and mobility using the Root et al[3,4] techniques. Clinicians should acknowledge poor reliability of first ray measurements, especially when making treatment decisions. In addition, further research is needed to determine the effects of force application and ankle position, (ie, subtalar and talocrural) on first ray reliability measurements. Finally, a validity study to compare the Root et al techniques with a gold standard is warranted.

ACKNOWLEDGMENTS

We thank HealthWorks Rehab & Fitness for the use of their facility. We also thank Dr S. Ayers for her time contribution in reviewing this manuscript. We thank all of our subjects; our examiners, Kevin Kotsko, Ray Adams, Suzanne Bologa, and Charis Mitchell; and our recorders, Scott Dietrich, Danielle Bifulco, Matt Wallace, Tim Kent, and Kim Samson, for all of their time and assistance in this research.

REFERENCES

1. Donatelli R, Wolf SL. *The Biomechanics of the Foot and Ankle.* Philadelphia, PA: FA Davis; 1990:23–26,141–142.
2. Glasoe WM, Yack HJ, Saltzman CL. Anatomy and biomechanics of the first ray. *Phys Ther.* 1999;79:854–859.
3. Root ML, Orien WP, Weed JH. *Normal and Abnormal Function of the Foot: Clinical Biomechanics.* Vol. II. Los Angeles, CA: Clinical Biomechanics Corp; 1977:48–51,265–266,285,344–345,350–354,358–359,363–367,376–377.
4. Root ML, Orien WP, Weed JH, Hughes RJ. *Biomechanical Examination of the Foot.* Los Angeles, CA: Clinical Biomechanics Corp; 1971:76–87.
5. Meyer JM, Tomeno B, Burdet A. Metatarsalgia due to insufficient support by the first ray. *Int Orthop.* 1981;5:193–201.
6. Glasoe WM, Allen MK, Saltzman CL. First ray dorsal mobility in relation to hallux valgus deformity and first intermetatarsal angle. *Foot Ankle Int.* 2001;22:98–101.
7. Klaue K, Hansen ST, Masquelet AC. Clinical, quantitative assessment of first tarsometatarsal mobility in the sagittal plane and its relation to hallux valgus deformity. *Foot Ankle Int.* 1994;15:9–13.
8. Bouysset M, Tebib J, Noel E, et al. Rheumatoid flat foot and deformity of the first ray. *J Rheumatol.* 2002;29:903–905.
9. Birke JA, Franks BD, Foto JG. First ray joint limitation, pressure, and ulceration of the first metatarsal head in diabetes mellitus. *Foot Ankle Int.* 1995;16:277–284.
10. Hamill J, Bates BT, Knutzen KM, Kirkpatrick GM. Relationship between selected static and dynamic lower extremity measures. *Clin Biomech.* 1989;4:217–225.
11. Glasoe WM, Yack J, Saltzman CL. Measuring first ray mobility with a new device. *Arch Phys Med Rehabil.* 1999;80:122–124.
12. Glasoe WM, Yack HJ, Saltzman CL. The reliability and validity of a first ray measurement device. *Foot Ankle Int.* 2000;21:240–246.
13. Glasoe WM, Allen MK, Saltzman CL, Ludewig PM, Sublett SH. Comparison of two methods used to assess first-ray mobility. *Foot Ankle Int.* 2002:23:248–252.
14. Cornwall MW, Fishco WD, McPoil TG, Lance CR, O'Donnell D, Hunt L. Reliability and validity of clinically assessing first-ray mobility of the foot. *J Am Podiatr Med Assoc.* 2004;94:470–476.
15. Cohen JA. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
16. Watson CJ, Propps M, Galt W, Redding A, Dobbs D. Reliability of McConnell's classification of patellar orientation in symptomatic and asymptomatic subjects. *J Orthop Sports Phys Ther.* 1999;29:378–385.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
18. Kinzey SJ. Significance, statistical or practical: the question of so what? *J Sport Rehabil.* 2000;9:267–268.
19. Hayes KW, Petersen CM. Reliability of classification derived from Cyriax's resisted testing in subjects with painful shoulders and knees. *J Orthop Sports Phys Ther.* 2003;33:235–246.
20. McPoil TH, Knecht HG, Schuit D. A survey of foot types in normal females between the ages of 18 and 30 years. *J Orthop Sports Phys Ther.* 1988;9:406–409.
21. Picciano AM, Rowlands MS, Worrell T. Reliability of open and closed kinetic chain subtalar joint neutral positions and navicular drop test. *J Orthop Sports Phys Ther.* 1993;18:553–558.
22. Elveru RA, Rothstein JM, Lamb RL. Goniometric reliability in a clinical setting: subtalar and ankle joint measurements. *Phys Ther.* 1988;68:672–677.
23. Bevans JS. The influence of subtalar joint position on first-ray dorsiflexion: a pilot study. *Br J Podiatry* 2003;6:69–72.
24. Grebing BR, Coughlin MJ. The effect of ankle position on the exam for first ray mobility. *Foot Ankle Int.* 2004;25:467–475.
25. Nawoczenski DA, Baumhauer JF, Umberger BR. Relationship between clinical measurements and motion of the first metatarsophalangeal joint during gait. *J Bone Joint Surg Am.* 1999;81:370–376.

COMMENTARY

Brian G. Ragan, PhD, ATC, CSCS

Editor's Note: Brian G. Ragan, PhD, ATC, CSCS, is an Assistant Professor in the Division of Athletic Training at the University of Northern Iowa, Cedar Falls, IA.

The issue of clinical measures and their usefulness is an important and relevant topic in athletic training. I am pleased to see work addressing validity evidence for clinical measures specifically for the foot and ankle. Clinicians use many techniques and measurements of the foot and ankle to evaluate and treat their athletes and patients. The authors have used an uncommon but appropriate criterion-referenced approach[1] in athletic training research to establish evidence of reliability for common foot and ankle measures of the first ray. The follow-

ing commentary is focused mainly on the measurement and statistical design issues of this study.

I am aware that this type of study and its methods have been used abundantly in the past to investigate intrarater and interrater reliabilities and the influence of experience on the reliability of a clinical measure.[2–5] I am concerned in general that we may be putting the cart before the horse in this case by first addressing sources of error such as experience. Although variation among raters is ultimately needed,[6] it seems more appropriate that the overall reliability or lack of reliability in the scores should be examined initially. An overall sense of the reliability of the person's characteristic or trait being measured is needed, with general reliability coefficients, before specifically investing effort to examine for possible sources of error (ie, experience). The measurement issue I have with this type of study is that the design does not match the intended purpose.

Although the approach has been used by many,[2–5] I do think there is a problem in answering the research question with the design. The methods in this study include a group of experienced (n = 2) and a group of inexperienced judges (n = 2) to rate and classify 36 people (2 feet per person, for a total of 72 feet) on 2 occasions. The research question compares the intrarater and interrater reliability of the experienced and inexperienced judges. On initial review, the sample size of 36 (72 feet) would appear to be adequate. The problem is that the sample size examining intrarater and interrater agreement with this design is only 4 examiners. The focus of the study is measuring a characteristic of the judges (agreement), and the focus must be on them as opposed to measuring the characteristic of the 36 subjects' first ray position and mobility. Ensuring that the design of the study focuses on the desired characteristic is vital. Currently, what conclusions can be made about the raters' characteristic experience with only 4 raters in the study?

The approach needed to answer this question would involve 2 groups of judges of a sufficient sample size to rate the same relatively small sample of feet (representing the distribution of foot types and motions would be ideal). This way, experience could be examined. This issue involving the number of judges needed for interrater reliability studies using norm-referenced standards, such as the intraclass correlation coefficient (ICC), has recently started to be addressed.[7] It has been suggested that the number of judges be equal to the number of subjects measured in reliability studies investigating interrater reliability when using ICCs.

I think this design issue is an important one. The authors in this study have followed a common design and method that are incorrect for answering the stated purpose of the study. My intent in the commentary is to aid in future investigations to avoid this design problem.

REFERENCES

1. Looney M. Criterion-referenced measurement: reliability. In: Safrit MJ, Wood TM, eds. *Measurement Concepts in Physical Education and Exercise Science.* Champaign, IL: Human Kinetics; 1989:137–152.
2. Bjorklund K, Skold C, Andersson L, Dalen N. Reliability of a criterion-based test of athletes with knee injuries; where the physiotherapist and the patient independently and simultaneously assess the patient's performance. *Knee Surg Sports Traumatol Arthrosc.* 2005 Jun 9; [Epub ahead of print].
3. Haight HJ, Dahm DL, Smith J, Krause DA. Measuring standing hindfoot alignment: reliability of goniometric and visual measurements. *Arch Phys Med Rehabil.* 2005;86:571–575.
4. Loughran S, Tennant N, Kishore A, Swan IR. Interobserver reliability in evaluating postural stability between clinicians and posturography. *Clin Otolaryngol.* 2005;30:255–257.
5. Shrader JA, Popovich JM Jr, Gracey GC, Danoff JV. Navicular drop measurement in people with rheumatoid arthritis: interrater and intrarater reliability. *Phys Ther.* 2005;85:656–664.
6. American Psychological Association, National Council on Measurement in Education, American Educational Research Association. *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association; 1999:ix, 1,194.
7. Saito Y, Sozu T, Hamada C, Yoshimura I. Effective number of subjects and number of raters for inter-rater reliability studies. *Stat Med.* 2005 Sep 5; [Epub ahead of print].

AUTHORS' RESPONSE

We thank Dr Brian G. Ragan for his review and appreciate his commentary on our article. We feel that this process will improve the methods used in future studies.

The first point that we would like to respond to is that regarding our comparison of experienced and inexperienced examiners. In our study, we examined overall intrarater reliability coefficients for both position and mobility measurements. In addition to that, we wanted to determine examiners' kappa and percentage agreement values for experienced and inexperienced examiners. We feel that we were able to successfully achieve 2 purposes in this paper with the number of examiners used.

The second point we would like to respond to is the suggestion of an increased number of examiners and/or decreased number of subjects. Walter et al[1] and Saito et al[2] suggested using designs in which the number of examiners and subjects is similar to minimize variance with intraclass correlation coefficients (ICCs). When the number of subjects exceeds the number of examiners, ICC variance increases.[2] We made an attempt to minimize interrater variance through the use of training sessions. We are unsure, though, how unequal numbers of raters and subjects affect kappa values. After reviewing the statistical calculation for kappa provided by Shoukri and Pause,[3] we postulate that the kappa value would decrease in response to an increase in the number of measurements per subject.

We did not conduct an a priori power analysis but determined our sample sizes based on sample sizes in published reliability studies as well as practicality. Sim and Wright[4] suggested that 2 examiners testing dichotomous variables with 25 to 35 subjects had sufficient power for detecting a kappa value of .50. Based on this work, we feel that the number of examiners and subjects was sufficient to achieve the desired power for our study. Although statistical significance was present, as mentioned in the ''Discussion'' section, we were more concerned with clinical meaningfulness than statistical significance.

Time, geographic constraints, and respondent burden often prevent a larger number of qualified examiners in a study.[1] Investigators may ask several potential examiners to participate before finding ones who are willing, as was the case in our study. Repetitive measurements by a large number of examiners can also fatigue subjects and can often take more time than they are willing to give.

Again, we appreciate Dr Ragan's comments and agree that future researchers should be aware of the issues of power and variance with unequal examiner and subject participants.

REFERENCES

1. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med.* 1998;17:101–110.
2. Saito Y, Sozu T, Hamada C, Yoshimura I. Effective number of subjects and number of raters for inter-rater reliability studies. *Stat Med.* 2005 Sep 5; [Epub ahead of print].
3. Shoukri MM, Pause CA. *Statistical Methods for the Health Sciences.* 2nd ed. Boca Raton, FL: CRC Press; 1999.
4. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther.* 2005;85:257–268.